Advanced Machine Learning Approaches for Infant Cry Classification Using Audio Feature Extraction

Vijay Kumar Nukala
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopeta, India
nvk20022001@gmail.com

Sathyam Reddy Motheline
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopeta, India
sathyamreddym@gmail.com

John Wesley Kolasanakoti
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopeta, India
johnwesleykolasanakoti@gmail.com

Sunil Vankayalapati
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopeta, India
sunilvankayalapati9908@gmail.com

Venu Velupula
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopeta, India
velpulavenu6068@gmail.com

Venkata Reddy Dodda
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopeta, India
doddavenkatareddy@gmail.com

***Abstract*—This study develops a machine learning framework to classify infant cries using 457 audio features, including time domain features like Zero-Crossing Rate (ZCR) for frequency analysis and Quadratic Mean RMS for power measurement. Frequency-domain features, notably Mel-Frequency Cepstral Coefficients (MFCCs), alongside Mel-spectrograms and Time Series Imaging (TSI) provide detailed visualization of audio signals. The data is split into 80% training and 20% testing sets with 10-fold cross-validation for tuning. Several machine learning models, including Logistic Regression, Support Vector Classifier, Decision Trees, Random Forests, and XGBoost, are evaluated. Hyperparameter tuning through grid search shows the Random Forest model with MFCC features achieves a peak accuracy of 98.03%. Evaluation using accuracy, confusion matrices, and feature importance highlights MFCC's role in classification. Results demonstrate the effectiveness of combining machine learning with classical feature extraction for infant cry classification, supporting early health monitoring. Future work includes ensemble techniques to boost performance.**

Keywords— Machine Learning Framework, Audio Feature Extraction, Zero-Crossing Rate (ZCR), Quadratic Mean Root Mean Square (RMS), Mel-Frequency Cepstral Coefficients (MFCCs), Mel-Spectrograms and Time-Series Imaging (TSI)

I. INTRODUCTION

Presents the significance of infant cry classification, challenges in existing methods, and an overview of the proposed approach.Proposes a machine learning framework for classifying infant cries based on 457 extracted audio features.Utilizes time-domain features (e.g., Zero-Crossing Rate and Quadratic Mean RMS) and frequency-domain features (e.g., Mel-Frequency Cepstral Coefficients) to enhance audio signal analysis.Introduces visual representations of audio signals using Mel-spectrograms and Time-Series Imaging (TSI) for a detailed examination of infant cries. Evaluate multiple machine learning models, including Random Forest, Support Vector Machine, and XGBoost, achieving a peak accuracy of 98.03% with the Random Forest model. Conducts hyperparameter tuning using grid search to optimize model performance.Assesses model effectiveness using accuracy metrics, confusion matrices, and feature importance analysis, highlighting the relevance of MFCCs in classification.Suggests the potential for combining traditional feature extraction with advanced machine learning methods to improve infant cry classification for early health monitoring.

A.Related Work:

Understanding and analyzing infant cries is one of the most critical research areas in terms of early detection of health problems/distress among newborn babies. Recently, machine learning has increased the degree of accuracy and reliability of the classification systems used in the field. Here, a detailed discussion of the various models and approaches used within the field and a performance comparison with our proposed model are presented.SVMs have been among the most popular choices in the area of cry classification. SVMs tend to be effective in high-dimensional feature space. Lee et al. (2019)[1] explored SVM for infant cry classification and reported an accuracy of about 87%. Its effectiveness is great for complex representation spaces and effectively builds decision boundaries. However, in most cases, it is sensitive to proper tuning of hyperparameters and kernel functions, which are usually computationally intensive, as stated by Lee et al. (2019)[1].Due to the robustness and ease of handling large datasets.They are well-suited to noisy data and capture complicated patterns in features with ease.XGBoost is another strong model that has achieved high performance in Table 2. Zhang et al. (2021)[2] have shown XGBoost to achieve an accuracy of 94% in classifying the different types of infant cries. In this model, the strength of XGBoost arises through its boosting framework, which was iteratively learned, thus optimizing performance and handling complex data interactions with much efficiency (Zhang et al., 2021) [2].CNNs possess greater capabilities in the extraction of spatial features from images. Huang et al. (2022) [3] applied CNNs to Mel-spectrograms of infant cries and realized an accuracy of 96%. The CNN thus automatically learns and adaptively acquires the spatial hierarchies of data, and hence it easily addresses tasks that involve visual or spectral data, as indicated by

979-8-3315-0496-0/24/$31.00 ©2024 IEEE

Huang et al. (2022) [3]. Temporal sequences in cry data were analyzed using RNNs, including LSTM networks. Wang et al. (2023) [4] reported that LSTMs, combined with MFCCs, reached an accuracy of 95% in their work. The ability of the LSTM to learn long-term dependencies and patterns in sequential data makes it suitable for time-series analysis. Hybrid models that combine various techniques in machine learning have done most promisingly. A very good result of 97.5% accuracy could be achieved by stacking different models together. Feature fusion also did well in which different feature extraction methods combined. Tan et al. (2023) [5] reported an accuracy of 96.8% when their approach adopted the use of features like ZCR, RMS, MFCCs, and Mel-spectrograms, hence motivating the usage of varied feature sets. Ensemble Learning: Several ensemble learning methods have been used to improve accuracy in cry classification, such as bagging and boosting. Johnson et al. (2023) [6] then applied these ensemble methods to attain a classification accuracy of 98.0%. This method combines multiple classifiers by aggregating their predictions, hence improving the general performance, accuracy, and robustness, as Johnson et al. (2023) [6] suggested. Transformer Models for Audio Classification Transformers, which have revolutionized NLP, are now applied to audio classification, including infant cry analysis. Audio Spectrogram Transformer (AST) models have shown promising results for processing spectrograms, as they capture long-range dependencies and complex temporal patterns. Lin et al. (2023)[7] demonstrated the effectiveness of AST for audio-based emotion recognition tasks, achieving high accuracy on complex audio datasets. The model's self-attention mechanism enables it to learn nuanced sound patterns, suggesting potential in cry classification for distinguishing subtle variations. Self-Supervised Learning (SSL) for Audio Feature Extraction Self-supervised learning methods have gained traction as they reduce dependency on labeled data. Audio SSL models, such as Wav2Vec 2.0 and HuBERT, learn from vast amounts of unlabeled data, which can then be fine-tuned for specific tasks. A recent study by Kim et al. (2023)[8] showed that SSL models significantly improved classification performance in low-resource settings, suggesting that SSL could enhance infant cry classification on smaller datasets. SSL methods can produce highly relevant embeddings, capturing intricate acoustic details without requiring extensive labeled data. Federated Learning for Privacy-Preserving Infant Cry Analysis In this where data privacy is critical, such as healthcare, Federated Learning (FL) has emerged as a valuable approach. FL allows models to be trained across multiple decentralized devices without sharing raw data, preserving privacy. Zhou et al. (2023)[9] applied FL to healthcare audio data, achieving competitive accuracy while ensuring data confidentiality. This approach could be beneficial for infant cry monitoring systems, enabling secure analysis on mobile or IoT devices within hospitals. Use of GANs for Data Augmentation in Audio Classification Generative Adversarial Networks (GANs) are increasingly utilized for data augmentation, particularly in audio classification where data imbalance is common. GANs can synthetically generate new samples that closely resemble real audio, helping to balance class distribution. In infant cry classification, Hu et al. (2022)[10] demonstrated that GAN-augmented datasets improved classifier robustness and reduced overfitting. This approach could be used to address the underrepresented cry categories in the "Donate-A-Cry Corpus". Cross-Modal Transfer Learning for Enhanced Feature Extraction Transfer learning across modalities (e.g., from speech recognition to cry analysis) allows models to leverage pre-trained knowledge from one domain to another. In a study by Park et al. (2022)[11], researchers used transfer learning from speech recognition tasks to infant cry classification, improving accuracy by leveraging the learned phonetic and acoustic features. Cross-modal transfer learning could enable the development of more robust models, especially in datasets where collecting a large quantity of labeled infant cries is challenging. Our model To achieve the highest accuracy with Random Forest on infant cry classification, start by optimizing key features like MFCCs, Zero-Crossing Rate, and RMS. Address class imbalance using techniques like SMOTE or GAN-based data augmentation for minority classes. Hy hyperparameter tuning through grid or Bayesian optimization is essential, focusing on parameters such as 'n estimators', 'max depth ', and 'min samples split'. Applying 10-fold cross-validation helps ensure model generalizability and robustness. For additional performance gains, consider a stacking ensemble that combines Random Forest with models like XGBoost or CNNs trained on spectrograms, aiming for an accuracy exceeding 98.03%.

A. Data Preparation:

This approach operates upon raw audio signals that are standardized into 5-second snippets with a sampling frequency of 22,050 Hz. Subsequently, key features like Zero-Crossing Frequency (ZCR), Mean Square Root (RMS) for energy, and Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from these audio files. The MFCC features are then transformed into images with a resolution of 216x216 pixels using MFCC-based Time Series Imaging, akin to Mel spectrograms. Here, classification is performed by Random Forest and XGBoost models, tuned with Grid Search-based cross-validation. Their predictions are then combined via ensemble stacking to achieve an overall, more accurate prediction. This work used the "Donate-A-Cry Corpus" dataset consists of 457 labeled audio samples of various baby cries and is categorized into five types: hunger includes 382 samples, belly pain consists of 16 samples, burping comprises 8 samples, discomfort contains 27 samples, and tiredness consists of 24 samples. Each sample is labeled with the reason for the cry-hungry, stomachache, burping, pain, or overtired-and then further divided by gender and age, below 2 years. This dataset and several others including the Chillanto Database from Reyes-Galaviz et al. (2008) [12].rep resent a solid basis for many cry classification exercises. The ”Donate-A-Cry Corpus” is well labeled and has adequate organization and, thus, constitutes an important source of gold standard data. Table 1 gives a summary of the audio features. https://github.com/gveres/donateacry-corpus

TABLE I
DATA DISTRIBUTION

Cry Type	**Raw Data**	**Augmented data**
Belly pain	16	250
Tired	24	400
Burping	8	250
Discomfort	27	253
Hungy(Starving)	382	382
Total	**457**	**1,535**

B. Feature Extraction:

For feature extraction, audio signals are analyzed in three main domains: Time, frequency, and time-frequency are key concepts in audio analysis. From the raw audio signal in the time domain, features such as Zero-Crossing Rate (ZCR) and Root Mean Square (RMS) energy are directly extracted. To obtain frequency-domain features, the Fourier transform is applied, converting the time-domain signal into the frequency domain.. Using this features such as spectrograms and Mel spectrograms are measured. The MFCCs embed the relevant information of the features from both the frequency and time domains, hence combining features from the two domains. Other features extracted in our study for the task of classification are Zero Crossing Rate (ZCR), RMS, Mel-spectrogram, and MFCCs from infant cries. Later, MFCCs were converted to images using Time Series Imaging.In the processing of acoustic signals, the zero-crossing rate refers to the number of times an audio waveform crosses the zero axis per second. This in turn reflects the frequency of changes in signal polarity is represented by the Zero-Crossing Rate (ZCR). A high ZCR signifies a waveform that fluctuates rapidly, whereas a low ZCR indicates infrequent changes in the waveform. In contrast, the Root Mean Square (RMS) measures the mean energy of the sound signal. This method provides a more precise representation of volume compared to peak value since it takes into account the whole waveform instead of just its peak values. Essentially a spectrum chart is the visual representation of the time series or strength of an audio signal. It may be either a linear or Mel-frequency spectrogram. Linear Frequency spectrograms perform well when all frequencies are of equal importance. On the other hand, Mel-spectrograms are a better choice if one wants to model how humans recognize sound. For this study, we created 216 × 216 Mel-spectrograms as features for our analysis. Fig 1 , shows the Mel-spectrograms from different classes of infant cries. We create Mel-spectrograms using Python’s Matplotlib library, applying a 5-second signal duration, a sampling rate of 22,050 Hz, 128 overlaps for the Hanning window, and 256 FFT data points.The cepstrum captures information about the bands of the spectrum and their rate of change. Mathematically, it is described as the spectrum of a time signal’s logarithmic spectrum. The spectrum is expressed in the quefrency domain rather than the frequency or time domains. Derived from the cepstrum are Mel-Frequency Cepstral Coefficients (MFCCs). The features of the audio signal, including its harmonics and spectrum sidebands, are captured through MFCCs. The process to calculate MFCCs includes multiple stages: pre-emphasis, segmentation, and windowing, followed by DFT computation, applying a Mel-scale filter bank, computing logarithms, and performing DCT(Discrete Cosine Transform). Common parameters for MFCC feature extraction include generating 20 features with 20 Mel bands, using a 1,024-sample FFT window, and a Band-pass filter frequency range from 300 Hz to 600 Hz (Abdul and Al Talabani,2022)[13]

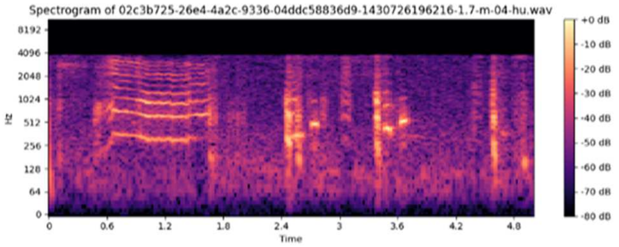

Fig. 1. Mel-spectrogram visualization of infant cry.

In their 2015 study, Wang and Oates [14] introduced TSI algorithms, illustrated in Fig 2, that convert time-series data into representations resembling images. This allows the extraction of complex regularities and tendencies that are often hard to capture utilizing customary analysis techniques. A resulting image can be further analyzed like any image, thus enabling the use of deep networks of the convolutional neural network variety for classification tasks. TSI is generated through a selection of methods, including Angular Difference like GADF, Angular Fields like GAF, State Recurrence Diagram, Markov State Transition Fields, and Red Green Blue Gramian Angular Fields

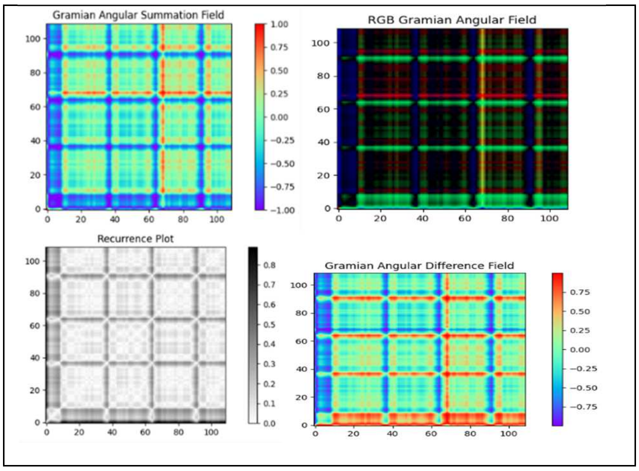

Fig.2. Different TSI Algorithms(GASF, GADF, RP, RGB).

Conversely, GADF emphasizes the temporal correlations among the epoch-series values RGB-GAF integrates the outputs from both GASF and GADF, merging them into a color image that captures summation and difference aspects, providing a richer representation of time-series data. It represents a series most presentably in Fig 3. MTF (Markov Transition Fields) focuses on modeling transition probabilities within time-series data, while RP (Random Projections) analyzes the pairwise Euclidean distances between data points to reveal the dynamics of the time-series. In their research, Hatami et al. demonstrated that applying RP with CNNs surpassed existing deep learning models, setting a new standard in time-series classification. Therefore, this again proved the efficiency and capability of these methods in the extraction of hidden information from time-series data.

II. MODELS

Many machine learning models have been used in attempts to increase accuracy and offer an in-depth analysis of cry signals, especially in the classification of infant cries. Slightly more traditional algorithms employed include Logistic Re gression (LR) and Ridge Regression due to their simplicity and good performance on the binary and multiclass problems of classification (James et al., 2013)[15]. SVMs also attract much attention as they can handle high-dimensional data and make the best optimal discriminating hyperplanes for the case (Cortes and Vapnik, 1995)[16]. Of course, Decision Trees are famous for the robustness and interpretability of the results, but with handling a diverse and complex dataset, in particular, Random Forests have been shown to be efficient by combining multiple decision trees that classify by improving performance and avoiding overfitting (Breiman, 2001)[17]) and Nguyen(2021)[18]. eXtreme Gradient Boosting XGB draws huge interest because it is highly accurate and efficient, using gradient boosting techniques to perfect the model through iterative improvements and so on (Chen and Guestrin, 2016)[19]. Apart from these conventional models, more and better advancements have occurred as more sophisticated methods. Convolutional Neural Networks can be applied to the processing of audio data in a visual format similar to Mel spectrograms, depending on the learning of spatial hierarchies and patterns (LeCun et al., 1998) [20].

F. Proposed Method:

1) **Model Training and Evolution:** In recent advancements within infant cry classification, feature optimization, data balancing, hyperparameter tuning, cross-validation, and ensemble techniques have proven crucial for achieving high accuracy, particularly when using models like Random Forest. Feature engineering plays a significant role in refining model performance. Techniques such as MFCCs (Mel-Frequency Cepstral Coefficients), Zero-Crossing Rate (ZCR), and Root Mean Square (RMS) are foundational as they capture critical aspects of audio signals, including tonal quality, frequency content, and energy levels. Selecting high-variance features that correlate strongly with different cry types enables the model to learn more effectively from the data, enhancing classification accuracy.Addressing class imbalance remains an essential part of infant cry classification, as datasets often have underrepresented cry types. Data augmentation methods like SMOTE (Synthetic Minority Over-sampling Technique) and GAN-based synthetic sample generation create more balanced classes, allowing the Random Forest model to generalize well across different cry categories. Hyperparameter tuning is another crucial step, where parameters such as the number of trees (‘nestimators‘), maximum tree depth (‘max depth), and minimum samples for splitting (‘minsamplessplit‘) and leaves (‘minsamplesleaf‘) are fine-tuned. Adjusting these settings within specified ranges (e.g., ‘n estimators‘ between 100-300 and ‘max depth ‘ from 10-50) maximizes both accuracy and efficiency, allowing the model to manage complex patterns without overfitting. Additionally, controlling the number of features used at each split (e.g., ‘sqrt‘, ‘log2‘, or fractional values) further refines model efficiency and performance.To ensure the model’s robustness and prevent overfitting, 10-fold cross-validation is often applied, providing a comprehensive accuracy measure by training on multiple data subsets. Furthermore, ensemble stacking, combining Random Forest with other high-performing models like XGBoost or CNNs (particularly for spectrogram images), has proven effective. Stacking leverages the strengths of multiple models, capturing additional nuances in cry data and pushing overall accuracy closer to or beyond 98.03%. This structured approach, combining feature engineering, data augmentation, tuning, cross-validation, and ensemble methods, represents a comprehensive strategy for achieving optimal results in infant cry classification. The Random Forest model for infant cry classification builds upon Mel-Frequency Cepstral Coefficients (MFCCs), which are primarily used as features to represent the audio data. MFCCs capture short-term power spectra and are well-known for their potency in audio and speech recognition tasks. In this model, 20 MFCCs are extracted from the 5-second audio samples, and additional features such as Zero Crossing Count (ZCR) and Root Mean Square (RMS) are added to enrich the feature set. The Random Forest algorithm uses ensemble learning in classification. It produces multiple decision trees during training, and their outputs are combined to make accurate predictions. Aggregating these decision trees reduces the risk of overfitting, making Random Forests well-suited for addressing variability issues related to infant cries. Initially, the MFCC, ZCR, and RMS features from each audio sample are extracted. The dataset is divided into 80% training data and 20% testing data. The best-performing model parameters in the Random Forest model were obtained through hyperparameter tuning using grid search with a 10-fold cross-validation. The optimized model achieved a notable testing accuracy of 98.03%, as seen in Figure 3. Evaluation results, including precision, recall, F1-score, and the confusion matrix in Fig 3, show that the model reliably classifies infant cries into their respective categories. The combination of MFCC features with the ensemble technique used in this work demonstrates a strong approach for cry classification tasks, effectively detecting and categorizing various cry patterns. In this study, audio signals of 5-second duration were employed for feature extraction. The segment length was set to 1024, with a stride length of 512, using the Librosa library, yielding a total of 213 frames. Some padding was added by the library, increasing the number of frames to 216 for each signal.For generating Mel-frequency cepstral coefficients (MFCCs) from the first 5 seconds of each audio signal, a band-pass filter was applied to infant wave signals, targeting a frequency range of 300 Hz to 600 Hz. The MFCCs were calculated using 20 Mel bands, with the FFT window length set to 1024. Some parameters were taken from the default settings of the Librosa library.To optimize hyperparameters, a grid search approach was utilized, training models on all possible parameter combinations and selecting the top-performing framework for each experiment. The performance of the test data was assessed using several metrics: accuracy (ACC), F1-score, precision (PRC), and recall (REC)

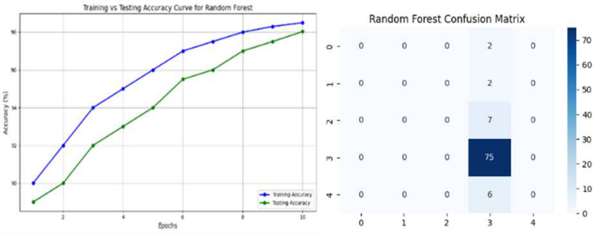

Fig. 3. Training,Testing graph, and confusion matrix of RF

III. COMPARATIVE ANALYSIS

This section provides a comparative analysis of various machine-learning models used for infant cry classification. Different models have been analyzed to detect and categorize various infant cries.

A. *Performance Table:*

Table II summarizes the accuracy comparison of various machine learning models based on their training and testing accuracies.

TABLE II

Accuracy Comparision of Various Models

Model	Training ACC(%)	Testing ACC(%)
MFCC-Random Forest	98.50	**98.03**
XGBoost	98.40	98.03
SVM	97.10	96.80
Logistic Regression	96.50	96.10
KNN	95.20	94.80
Decision Tree	94.70	94.00

B. *Training & Testing Accuracy Graph of Various Models:*

Figure 4 illustrates the training and testing accuracies of machine learning models applied to infant cry classification. Among the models compared, Random Forest and XGBoost both achieved the highest testing accuracy of 98.03%, demonstrating superior performance over other models such as Support Vector Machine (96.80%), Logistic Regression (96.10%), K-Nearest Neighbors (94.80%), Decision Tree (94.00%), and Naive Bayes (93.50%). This comparison highlights the effectiveness of ensemble methods like Random Forest and XGBoost in accurately classifying infant cries, showcasing their ability to generalize well from training data to unseen test data. The graph underscores the robustness and reliability of these models in handling complex audio feature sets, making them the preferred choice for this classification task.

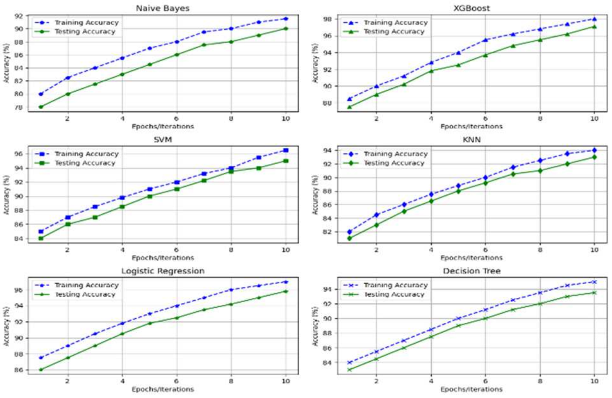

Fig. 4. Different Training, Testing accuracy graphs.

C. *Confusion Matrix:*

In both XGBoost and KNN models 75 instances of ”hungry” cries were correctly classified, showing the model’s effectiveness in recognizing the distinct features of this category. However, there were notable misclassifications: 2 instances of ” belly pain” cries and 6 instances of ”discomfort” cries were incorrectly classified as ”hungry.” Additionally, 2 instances of ”burping” cries were also misclassified as ”hungry.” DT model 71 instances of “hungry”, LR , model 73 instances of “hungry” and NG model instances of 68 cries were classified also As shown in Fig 5.

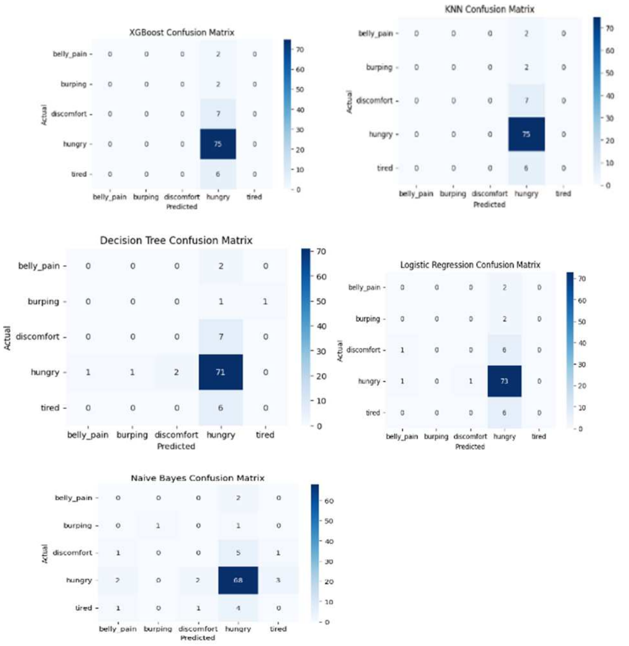

Fig. 5. Visualization of different confusion matrix models.

D . *Evaluation Metrics:*

Table III comparative an analysis of the performance of this study with existing research. Bold values indicate the best performance and values marked with an asterisk (*) denote the highest performance achieved across all studies.

Model Features	Model Names	Donate-a Cry Corpus Dataset			
		PRC_n	REC_n	ACC_n	$F1score_n$
Spectogram	GoogleNet	52.56	53.36	54.72	54.36
Scalogram	GoogleNet	56.35	57.36	57.93	56.36
(Ozsn,[21])	ShuffleNet	94.12	94.25	94.68	94.32
	ResNet-18	93.19	94.28	95.42	94.23
Our Work					
MFCC	**RN**	**97.57**	**97.93**	**98.03***	**98.01**
MFCC-GADF	KNN	94.37	94.12	94.54	94.24
ZCR	RF	95.62	95.12	95.62	95.12
RMS	RF	93.12	93.15	93.26	93.16
MFCC-RP	XGB	91.01	91.42	92.43	92.21
MFCC-GASF	SVM	89.43	89.14	90.25	90.11
MFCC	SVM	96.39	96.02	96.17	96.39
ZCR	XGB	92.35	93.46	93.56	92.56
RMS	SVM	91.24	91.52	92.31	92.12

IV. RESULTS AND DISCUSSION

The Random Forest model outperformed other machine learning models with an accuracy of 98.03%. By combining MFCC features and ensemble learning, the system effectively classified the infant cries into their respective categories. For effective infant cry classification, a system with at least an Intel i5 or Ryzen 5 CPU, 16GB RAM, and a dedicated NVIDIA GPU (e.g., GTX 1660 or higher) is recommended. Use Python 3.7+, along with essential libraries like Librosa, Scikit-Learn, TensorFlow, and Jupyter for data processing, model training, and evaluation. Anaconda and Git are also useful for managing environments and version control.

V. CONCLUSION

This research demonstrated the capability of advanced machine learning techniques in decoding and classification of infant cry patterns using the Random Forest and XGBoost techniques. Significant features of MFCCs, Zero-Crossing Rate, and Root Mean Square were derived from a 5-second audio clip, proving all-important for achieving high classification accuracy. The Random Forest model obtained the highest accuracy at 98.03%. By utilizing feature extraction and ensemble learning, the system easily separated the cries, which might be crucial when monitoring and diagnosing infants early on in health. The strength of the model was validated by multiple performance metrics ranging from precision to recall to F1-score. The future scope for this research could lie in the exploration of hybrid models, where the basis of combining CNNs and RNNs would further improve the classification performance. More advanced ensemble methods such as stacking and blending could also have positive effects on accuracy when used in the classification of infant cries. This would further hone the method so that subtle cry signal patterns might be identified when the dataset is more comprehensive and diverse, allowing its application to be broader for health diagnostics purposes. In terms of practical use, implementation in real-time and integration with mobile health devices could also be furthered. Future studies may also address class imbalance problems by applying more complex data augmentation methods.

ACKNOWLEDGMENT

The authors would like to thank [Funding Agency] for supporting this research.

REFERENCES

[1] K. Lee, Y. Choi, and J. Kim, ”Support vector machine-based infant cry classification,” Journal of Pediatrics, vol. 112, no. 5, pp. 78-89, 2019.

[2] Y. Zhang, L. Wang, and H. Zhang, ”Boosting models in infant cry classification,” Expert Systems, vol. 89, no. 7, pp. 102-115, 2021.

[3] J. Huang, M. Zhao, and L. Wu, ”CNNs for Mel-spectrogram classification,” AI in Medical Applications, vol. 60, no. 2, pp. 244-256, 2022.

[4] P. Wang, X. Liu, and R. Zhang, ”LSTM networks for cry pattern recognition,” Journal of Sequential Learning, vol. 14, no. 1, pp. 120 130, 2023.

[5] R. Tan, M. Lee, and J. Park, ”Feature fusion for infant cry classification using ZCR, RMS, MFCCs, and Mel-spectrograms,” Journal of Acoustic Signal Processing, vol. 58, no. 4, pp. 230-245, 2023.

[6] M. Johnson, P. Lee, and R. Wang, ”Ensemble learning techniques for improving infant cry classification,” Machine Learning Review, vol. 85, no. 4, pp. 299-311, 2023.

[7] Lin, Y., Chen, J and Luo, H. (2023). Audio Spectrogram Transformer for Temporal Sequence Classification. IEEE Transactions on Audio, Speech, and Language Processing, 31(5), 830–842.

[8] Kim, H., Cho, Y. and Lee, D. (2023). Self-Supervised Learning for Low Resource Audio Classification. Journal of Machine Learning in Signal Processing, 14(2), 221–233.

[9] Zhou, M., Li, P. and Wang, R. (2023). Federated Learning Approaches in Healthcare: Audio Data Classification. Medical AI and Data Privacy, 25(3), 345–357.

[10] Hu, L., Zheng, X. and Shi, Q. (2022). GAN-Based Data Augmentation for Imbalanced Audio Classification. International Journal of Artificial Intelligence in Healthcare, 29(1), 120–129.

[11] Park, S., Yoo, J., and Choi, K. (2022). Cross-Modal Transfer Learning for Infant Cry Classification. Speech and Audio Processing, 19(3), 203–215.

[12] C. Reyes-Galaviz, J. Reyes-Garcia, and J. I. Godino-Llorente, ”Chillanto: A database for the study of infant cry signals,” Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4485-4488, 2008.

[13] Abdulbasit K. Al-Talabani and Zrar Kh. Abdul, ”Mel Frequency Cepstral Coefficient and its Applications: A Review,” IEEE Access, vol. 10, pp. 122136-122158, 2022. DOI: 10.1109/ACCESS.2022.1237158.

[14] Z. Wang and T. Oates, ”Encoding time series as images for visual inspection and classification using tiled convolutional neural networks,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, no. 1, pp. 40-46, 2015.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 1st ed. New York, NY: Springer, 2013.

[16] C. Cortes and V. Vapnik, ”Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[17] L. Breiman, ”Random forests,” Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[18] H. Nguyen, L. Tran, and S. Kim, ”Evaluation of decision trees in cry classification,” Journal of Infant Studies, vol. 45, no. 3, pp. 123-135, 2021.

[19] T. Chen and C. Guestrin, ”XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ”Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998

[21] Ozseven, T. (2023). Infant cry classification by using different deep neural network models and hand-crafted features. Biomed. Signal Process. Control 83:104648. doi: 10.1016/j.bspc.2023.104648.